Analyzing Tagged Resources for Social Interests Detection

Manel Mezghani[1,2], André Péninou[2], Corinne Amel Zayani[1], Ikram Amous[1] and Florence Sèdes[2]
[1]*Department of Computer Science, Sfax University, MIRACL Laboratory, Sfax, Tunisia*
[2]*Department of Computer Science, Paul Sabatier University, IRIT Laboratory, Toulouse, France*
mezghani.manel@gmail.com, peninou@irit.fr, {corinne.zayani, ikram.amous}@isecs.rnu.tn, sedes@irit.fr

Keywords: User Interests, Tagging Behaviour, Resources, Social Network, Adaptation.

Abstract: The social user is characterized by his social activity like sharing information, making relationships, etc. With the evolution of social content, the user needs more accurate information that reflects his interests. We focus on analyzing user's interests which are key elements for improving adaptation (recommendation, personalization, etc.). In this article, we are interested to overcome issues that influence the quality of adaptation in social networks, such as the accuracy of user's interests. The originality of our approach is the proposal of a new technique of user's interests detection by analyzing the accuracy of the tagging behaviour of the users in order to figure out the tags which really reflect the resources content. We focus on semi-structured data (resources), since they provide more comprehensible information. Our approach has been tested and evaluated in *Delicious* social database. A comparison between our approach and classical tag-based approach shows that our approach performs better.

1 INTRODUCTION

The social user is characterized by his social activity like sharing information, making relationships, etc. With the evolution of social content, the user needs more accurate information that reflects his interests in order to provide better adaptation. A profile which reflects the appropriate characteristics (interests, preferences, etc.) could avoid cognitive overload and disorientation of the user when accessing the information space.

The adaptation is a process strongly related to the user modelling. In fact, each user has specific needs and then he needs specific adaptation. We are interested in adaptation of semi-structured data (resources) like the work of (Zayani et al., 2007) which proposes a priori adaptation of semi-structured data independently of the domain of application. We also want to address the work of (Rebai et al., 2013) which proposes an adaptation of navigation method based on semi-structured data by analyzing the navigation behaviour of the user. In our work, we analyze the tagging behaviour of each user applied to a semi-structured data (resources) in order to extract relevant interests. These interests could be used for an adaptation purpose in further works.

User's interests are extracted in a popular way, from his own profile (e.g. interests attribute) or from his social behaviour (e.g. tagging behaviour) or his social network (e.g. friends). However, detecting user's interests is a crucial problem (Milicevic et al., 2010). In fact, the user profile building process suffers from the lack of information provided by the user. Indeed, the user generally doesn't give all the information related to his interests and then the user profile can never be considered fully known by a system. So, in order to overcome such a problem, the researchers have analyzed the social environment of the user (Astrain et al., 2010) such as his neighbours (persons connected to the user explicitly or implicitly), his tagging behaviour (the action of tagging resources), or even the objects (resources) he interacts with.

Through an architecture of adaptation proposed in (Mezghani et al., 2012a), we will propose an interests detection approach. The originality of our approach is the proposal of a new technique of user's interests detection by analyzing the accuracy of the tagging behaviour of the user in order to figure out the tags which really reflect the resources content.

In this paper, we first show how researches have integrated the tag information, the neighbours and the resources content to detect user's interests for an adaptation purpose. In the second section, we propose our method of detecting interests and the experiments done to validate it. Finally, we conclude and propose our future works.

2 STATE OF THE ART

In this section, we discuss researches done in order to detect user's interests in social context. Then we discuss the main differences between our approach and the other researches.

2.1 Detecting User's Interests

Detecting user's interests exists before the creation of social networks. It fact, researchers analyze the user behaviour by observing his actions on a web page like time spending reading the web page, number of the movements of scrolling, etc., or by extracting keywords from the document already read.

As these classic solutions aren't always efficient to reflect the real user's interests (Ma et al., 2011) and with the creation of social networks, new solutions are envisaged. In fact, the user becomes an active contributor for creating social content and then new users manners are created which contribute to deduce his interests. According to (Astrain et al., 2010), interests could be deduced based on the user, the object or even the tag. The tagging behaviour is described as the connection of these three elements, since it is described as the action of tagging a resource (object) by a user. We discuss some researches done based on each element.

For the user, social interests are detected based on the user's profile. The profile could be explicitly provided by the user profile (Zayani et al., 2007), or implicitly deduced from his behaviour of navigation (Rebai et al., 2013) or behaviour of tagging. The user-based interests detection is also related to the other users in the network (neighbours). Even that neighbours could disorient the user (e.g. as spammers), they could be used for detecting pertinent interests (Kim et al., 2011). Social interests could also be deduced from users' relationships like (Tchuente et al., 2013) who extract social interests from the egocentric network of the user.

For the object, interests are deduced based on the objects that the user access. (White et al., 2009) combine interests of other users who have visited the same web page in order to recommend web pages. More recently, (Ma et al., 2011) combine the interests from different sources with a semantic reasoning, in order to extend the interests list. Objects could be any type of resource (text, image, etc.). Even that these works consider the resource, they do not analyze their content. To analyze the resource content, different techniques exist such as the indexation technique. This latter, is used in order to extract the significant terms from resources. After indexing resources, different scoring function could be applied in order to detect the most relevant resource according to a specific query (Vallet et al., 2010).

For the tag, it has proved its utility to detect user's interests (Kim et al., 2011). The tag is a user-generated keyword which reflects the users opinion in a resource (Astrain et al., 2010). Tag-based interests detection is based on the user's tagging behaviour by: i) integrating the history of the tagging behaviour in order to recommend tags or content e.g. (Wang et al., 2010), ii) analyzing used tags (Meo et al., 2010), iii) combining user's interests with information extracted from his tagging activities in order to recommend tags (Godoy and Amandi, 2008) or even iv) use the tag information by analyzing the semantic of tags (Kim et al., 2011).

2.2 Synthesis

After presenting some researches done to analyse the tagging behaviour elements, we discuss the main differences between our approach and the other researches.

Unlike most of the researches which focus on the tag content considered as an interest (by analyzing the semantic of the tags for example), we focus on analyzing the accuracy of the tags with the resources content.

We focus on analyzing the object-based rather than the user-based interests detection. In fact object-based interests detection provides richer information than the user-based method (Song et al., 2011).

For object-based interests detection, most of researches do not consider the accuracy of the tags with the object (resource) content. This problem has been cited in (Milicevic et al., 2010). However, the proposed approaches use techniques such as clustering, semantic treatment, etc. and neither of them analyze the resources content in their works.

Dealing with the accuracy of the tag could overcome problems related to the nature of these social annotations. The first problem is that these tags could be considered as personal and reflect the "feeling" of the user and rather than the content of the resource. Example: "good", "awesome", etc. The second problem is the ambiguity associated with the tags since they are user generated keywords and do not follow any rules. These problems have been treated explicitly in some researches (see (Mezghani et al., 2012b) and (Milicevic et al., 2010) for more details). In our approach, these drawbacks will be treated implicitly while detecting the accurate tags.

To summarize, our approach uses tags and treats them according to the content of their respective re-

sources. Accurate tags are those reflecting the resources content. The result is validated by matching accurate tags found by our approach (which uses user's neighbours) with the user's real tags (those used by the user). This validation method has been proposed by (Tchuente et al., 2013).

3 APPROACH FOR DETECTING USER'S INTERESTS

In this section, we use the architecture of the social adaptation proposed in (Mezghani et al., 2012a). We focus on the modules used from this architecture for developing the user's interests detection approach. Then, we propose our approach for detecting the user's interests.

3.1 Architecture

User's interests detection is a part of the user modelling module extracted from the proposed architecture of adaptation of social navigation in (Mezghani et al., 2012a) (see figure 1). The adaptation of social navigation is reached through a recommendation technique, which needs pertinent information about user's interests.

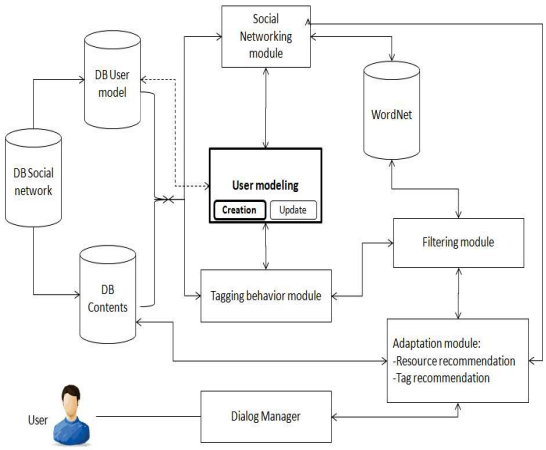

Figure 1: Architecture of adaptation of social navigation (Mezghani et al., 2012a).

In order to achieve our goal, we use the databases:

- The *DB Social Network*: This database contains data about the objects in the social network including the data about the resources and users. The data exploited in this approach are extracted from a specific social network (*Delicious*, *CiteULike*, *Last.fm*, *movieLens*, etc.). Social information adapted is depending on the social network (e.g. bookmarks in *Delicious*, scientific articles in *CiteUlike*, music in *Last.fm* and video in *movieLens*).
- The *DB User Model*: This database uses information from the *DB social network*. This module specifies information about users and networks of users (interests, preferences, friends, professional relationships, etc.).
- The *DB Contents*: This database uses also information from the *DB social network*. This module stores information about the resources of the social network (type of resource, tags associated by each users, metadata, etc.).

We detail the essential modules for detecting user's interests such as the *social networking module* and the *tagging behaviour module* as follows:

- *Social Networking Module*: This module exploits the *user modelling* module by analyzing the similarity between users to build networks of similar users using same tags and access the user's profiles to build networks of friends. This module is able to identify similar users with a similar tagging behaviour. Based on social relation, it is able to send information such as most popular users, friends, etc. for the *adaptation module*. So, the user's neighbours are extracted from *Social networking module*.
- *Tagging Behaviour Module*: This module contains data about the users who tags the resources of various types (e.g. photos, videos, scientific papers, etc.). Generally this activity is represented in a tripartite model (see equation 1) which describes the users U, the resources being tagged R and the tags T.

$$Tagging\_relation :< U, T, R > \quad (1)$$

3.2 The Development of the User Modelling Module

This module aims to create and update the user profile. We focus on the creation process (through the *creation* sub-module) which includes the interests detection approach. In our approach, we analyse the tags assigned to the resources to detect user's interests. This approach of detecting interests is applied to each user's neighbours. The neighbours could be detected through different ways (an egocentric network, users in the same community, users sharing common behaviours, etc.).

Before explaining our approach, we focus on the preparation of the data used to achieve an efficient interests detection approach (figure 2).

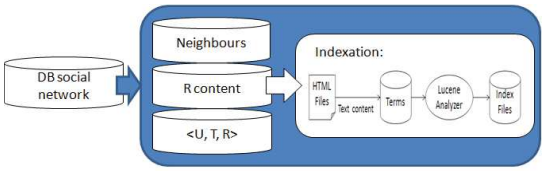

Figure 2: Data preparation.

We extract in the first step the data used in our approach: i) the tagging behaviour relations ($< U, T, R >$), which are composed of the tags applied to the resources by each user. This information is extracted from the *tagging behaviour module*. ii) The resources content which is extracted from the *DB content model* and iii) the neighbours which is extracted from the *DB user model*.

In the second step, we index the resources, using the *Lucene* API [1]. *Lucene* is a tool for indexing and searching technology. We use it in order to figure out the tags which are the most accurate with regards to the content of the tagged resource. *Lucene* is a field-based indexation technique. This characteristic allows indexing the documents according to one or more fields. For example, fields could be the title, the content, the URL, etc. In our approach, the indexation process has been made according to the content of the resource. The indexing process is explained as follows: when *Lucene* indexes documents, it divides them into a number of terms. Then, it stores the terms in an index file, where each term is associated with the document content. Terms are generated using an analyzer that converts each word in its root form. When a request is made it is treated by the same analyzer used to build the index and then used to find the corresponding term(s) in the index. This provides a list of documents matching the query.

After preparing the data, we explain the process of detecting the user's interests. The following steps are iterated for all tags of each user's neighbours.

In the first step, we generate the resources relevant to a specific query. A query is considered as a tag. This step is performed according to the index file generated from the indexation process.

In the second step, we assign a score to each relevant resource (issued from the previous step) according to the assigned tags. This score is the result of a function of similarity which takes into consideration the resource (as a semi-structured resource) and the query (as a tag). Many similarity functions exist in the literature such as the similarity function supported by *Lucene* [2]. We choose a predefined function of similarity which is a variant of the TF-IDF scoring model. The choice of such a model is due to the fact that the TF-IDF is an efficient and simple algorithm for matching words in a query to resources that are relevant to that query. However, the main limitation of such a model is that it doesn't take into consideration the relations between words (e.g. synonyms). The scoring function provides a result of the top-k resources relevant to the query q (the tag). In the last step, we test if the resource tagged by the query q exists in the top-k resources provided by the scoring function. If it is the case, we state the tag as relevant to the resource.

In order to validate our finding, we compare the founded relevant tags (our approach applied to user's neighbours) with the user's tags (real tagging behaviour). The validation of the detected interests, is done through a test of existence of the detected interests (from the neighbours) with the user's interests. This test is done through two methods:

- By a simple matching technique (i.e. if user-tag="picture" and neighbour-tag="picture", then the tag" picture" is considered relevant).
- By taking into consideration the synonyms and the related words (i.e.: if user-tag="pictures" or "photo" and neighbour-tag="picture", then the tag "picture" is considered relevant). The synonyms and related words are detected by interrogating *Wordnet* [3].

4 EXPERIMENTATION

We experiment our approach through the *Delicious* social database. First, we evaluate our approach according to the community of the user (issued from a specific community detection algorithm). The evaluation aims to compare the founded results with the existing user's tags. The comparison could be done by a simple matching technique or by considering the synonyms and related words. We test these two methods, and we retain the one which provides the best results to do the rest of the evaluations. Finally, we compare our approach (that analyzes the tagged resource) with the tag-based approach that uses the tag information without any pre-treatment (tags provided directly from the user).

4.1 Evaluation According to the Community

The definition of the term community varies from a

[1] http://lucene.apache.org/

[2] See http://lucene.apache.org/core/3_6_2 /api/core/org/apache/lucene/search/Similarity.html

[3] http://wordnet.princeton.edu/

work to another. In our work, we use the definition proposed by (Cazabet et al., 2010) and used in (Tchuente et al., 2013). The community is detected through an algorithm called "iLCD" which has proven his utility. We use this algorithm in order to generate communities associated to the database. The *Delicious* database contains social networking, bookmarking, and tagging information. This database is extracted from (Ivan et al., 2011). We present some statistics of the data present in this database: 1867 users, 69226 URLs and 53388 tags.

We run our approach on all the users of the database. These users have different number of neighbours (which may vary from 1 to 50 neighbours). The number of tags, resources and tagging relations is different for each user. This number may roughly vary from 3 to 800 for the tags, from 10 to 450 for the resources, and from 20 to 500 for the tagging relations.

We calculate the precision of the detected interests according to the tags in the neighbours' profiles. The precision P(u) for each user u is calculated according to the number of accurate tags ($C_u \subseteq I_u$) and the total number of tags provided as accurate (I_u) (formula 2).

$$P(u) = |C_u|/|I_u| \quad (2)$$

Our approach has been tested with different value of k such as k=20, k=50 and k=100. We calculate the average precision for all users (formula 3) provided from the precision formula P(u) for the user u. Where n=number of the users (in our case n=1867).

$$Average\_Precision = \sum_{i=1}^{n} P(u)/n \quad (3)$$

We calculate the precision for both methods of evaluation (the matching technique and with the synonyms and related words) (figure 3).

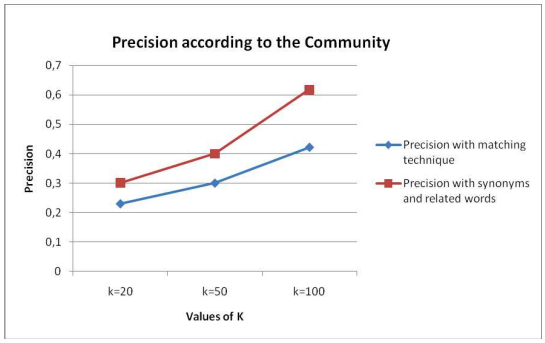

Figure 3: The average precision according to k=20, k=50 and k=100 according to the community.

We clearly see that the average precision that takes into consideration synonyms and related words is better that the matching technique. This is obvious because users may have the same interests but they may describe them differently.

We choose k=100 for the rest of the evaluation since it provides better results. We calculate the average precision of all users in the database according to i) the matching technique and according to ii)synonyms and related words. The average precision for the matching technique is 42.15% and for synonyms and related words is 61.85%. We notice that the consideration of synonyms and related words provides higher precision values. Moreover, we notice that the precision (for the two methods of comparison) varies according to different cases: i) the precision is higher for active users (having a lot of neighbours and a lot of tagging behaviour). ii) the precision is less higher for less active users. iii) the precision is equal to zero for some users due to the incoherence of the amount of information provided by the user versus his neighbours or vice versa.

We test if our approach treated the ambiguity of the tags. We notice that, the accurate interests provided by our approach are comprehensible keywords which reflect really the resource content like "technology", "foursquare", "history", etc. This is an advantage since tags are user-generated keywords. Our approach has filtered ambiguous tags (e.g.:"gis") that are not comprehensible by other users. Tags ambiguity has decreased from 35% to 10% according to WordNet, without developing an explicit method for tag filtering.

4.2 Comparison Between Our Approach and the Tag-based Approach

Using the same set of users, we compare our approach with the tag-based approach that uses only the tag information. We compare according to the k=100 of our approach (since it provides better results). Also, we compare by taking into consideration synonyms and related words since it is better than the matching technique. We calculate the average precision of all users in the database and compare it with the average precision provided by our approach. The precision of our approach is equal to 61.85% and of the tag-based approach is equal to 32.55%.

Our approach overcomes the tag-based approach in term of precision. This is due to the consideration of the content of the resources analyzed for the selection of relevant tags. The selection process has implicitly filter ambiguous tags that may not be comprehensible for other users. Consequently, we obtain a higher precision that the tag-based approach.

5 CONCLUSION

In this paper, we have proposed a new approach of detecting social interests. This approach is based on analyzing the tagging behaviour of each user. In fact, the analyze aims to extract the most accurate tags according to their relevance to the content of the tagged resources. We have tested our approach in *Delicious* database. The validation of the results is done by comparing the tags of each user with the result of tags issued from our approach using users' neighbours (his community). The proposed approach is able to detect potential user's interests by analyzing their social behaviour. Moreover, it improves the quality of the detected interests (tags) by decreasing implicitly their ambiguity. Then, it could be used for a purpose of adaptation since it provides a solution for detecting users' interests.

In future works, we will validate our approach according to another type of neighbours such as the egocentric network. This test allows us to deduce the neighbours which reflect the most the user's interests.

REFERENCES

Astrain, J. J., Cordoba, A., Echarte, F., and Villadangos, J. (2010). An algorithm for the improvement of tag-based social interest discovery. pages 49–54.

Cazabet, R., Amblard, F., and Hanachi, C. (2010). Detection of overlapping communities in dynamical social networks. In *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pages 309–314.

Godoy, D. and Amandi, A. (2008). Hybrid content and tag-based profiles for recommendation in collaborative tagging systems. In *Latin American Web Conference, 2008. LA-WEB '08.*, pages 58–65.

Ivan, C., Peter, B., and Tsvi, K. (2011). *HetRec '11: Proceedings of the 2Nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. ACM, New York, NY, USA.

Kim, H.-N., Alkhaldi, A., El Saddik, A., and Jo, G.-S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications*, 38(7):8488–8496.

Ma, Y., Zeng, Y., Ren, X., and Zhong, N. (2011). User interests modeling based on multi-source personal information fusion and semantic reasoning. In *Proceedings of the 7th International Conference on Active Media Technology*, AMT'11, page 195205, Berlin, Heidelberg. Springer-Verlag.

Meo, P. D., Quattrone, G., and Ursino, D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, 20(1):41–86.

Mezghani, M., Zayani, C. A., Amous, I., and Gargouri, F. (2012a). An extended architecture for adaptation of social navigation. In Krempels, K.-H. and Cordeiro, J., editors, *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, 18 - 21 April, 2012*, pages 540–545. SciTePress.

Mezghani, M., Zayani, C. A., Amous, I., and Gargouri, F. (2012b). A user profile modelling using social annotations: A survey. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, page 969976, New York, NY, USA. ACM.

Milicevic, A. K., Nanopoulos, A., and Ivanovic, M. (2010). Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, 33(3):187–209.

Rebai, R. Z., Zayani, C. A., and Amous, I. (2013). An adaptive navigation method for semi-structured data. In Morzy, T., Hrder, T., and Wrembel, R., editors, *Advances in Databases and Information Systems*, number 186 in Advances in Intelligent Systems and Computing, pages 207–215. Springer Berlin Heidelberg.

Song, Y., Zhang, L., and Giles, C. L. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Trans. Web*, 5(1):4:14:31.

Tchuente, D., Canut, M.-F., Jessel, N., Peninou, A., and Sdes, F. (2013). A community-based algorithm for deriving users profiles from egocentrics networks: experiment on facebook and DBLP. *Social Network Analysis and Mining*, 3(3):667–683.

Vallet, D., Cantador, I., and Jose, J. M. (2010). Personalizing web search with folksonomy-based user and document profiles. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rger, S., and Rijsbergen, K. v., editors, *Advances in Information Retrieval*, number 5993 in Lecture Notes in Computer Science, pages 420–431. Springer Berlin Heidelberg.

White, R. W., Bailey, P., and Chen, L. (2009). Predicting user interests from contextual information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 363370, New York, NY, USA. ACM.

Zayani, C. A., Pninou, A., Canut, C. M.-F., and Sedes, F. (2007). Towards an adaptation of semi-structured document querying. In Doan, B.-L., Jose, J. M., and Melucci, M., editors, *Proceedings of the CIR'07 Workshop on Context-Based Information Retrieval in conjunction with CONTEXT-07, Roskilde, Denmark, 20 August 2007*, volume 326 of *CEUR Workshop Proceedings*. CEUR-WS.org.